View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

# Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus.

Osvaldo Spositto, Gabriel Blanco, Lorena Matteo, Marcelo Levi, Julio Bossero.
Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza. Florencio Varela 1902, San Justo, Prov. Buenos Aires, Argentina
{sposito, g2blanco, lmatteo,mlevi,jbossero}@unlam.edu.ar

## RESUMEN

La Diferencia Esperada entre Progenies (DEP) es un indicador numérico que predice la calidad genética de las futuras crías de un toro o una vaca respecto de una base de comparación. Este es un valor genético que proporciona la mejor manera de comparar reproductores por la producción esperada en sus descendencias. En este trabajo estudiamos el uso de técnicas de Minería de Datos, del tipo Supervisadas y No Supervisadas, para identificar patrones o grupos de características en los valores genéticos de los animales reproductores que determinen el peso de los terneros al nacer de la cría de la raza Aberdeen Angus. El objetivo es brindar una herramienta complementaria para que un criador ganadero pueda seleccionar mejor los reproductores que al ser apareados con sus vientres, produzcan progenies superiores. En estos primeros estudios emplearon datos provenientes de 360 animales proporcionados por un establecimiento ganadero de la provincia de Buenos Aires. Se espera que a partir de los modelos aprendidos por los algoritmos se pueda extraer información preliminar sobre el valor genético de un animal, que pueda resultar de gran utilidad en el sector ganadero en la toma de decisiones en un programa de mejoramiento genético.

**Palabras clave:** Diferencia Esperada entre Progenies, Minería de Datos, Algoritmos Supervisados y No Supervisados, Mejoramiento Genético.

## CONTEXTO

La línea de investigación aquí presentada está enmarcada dentro del Programa de Incentivos para Docentes Investigadores de la Secretaría de Políticas Universitarias (PROINCE) 2019-2020. El mismo lleva el título: *Uso de Minería de Datos para Mejoramiento Genético en la raza Aberdeen Angus.* Este proyecto es financiado por la Universidad Nacional de La Matanza (UNLaM). Los datos utilizados para este estudio provienen de los rodeos Aberdeen Angus de la estancia El Doce y de Cabaña Las Lilas[1] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires.

Esta Cabaña perteneciente a la Asociación Argentina de Angus[2], es uno de los caudales genéticos más importantes de la Argentina y del Mercosur.

## 1. INTRODUCCIÓN

El propósito de un programa de mejoramiento genético de una raza de carne es conocer y promover los mejores animales basados en registros de comportamiento y evaluación de sus progenitores [1]. Los productores ganaderos se basan en ellos para identificar y procurar aquellos animales que mejor se adapten a las condiciones de producción existentes y que al mismo tiempo conduzcan a un incremento del beneficio económico de la actividad. Para esto es necesario valerse de información objetiva y precisa sobre los reproductores, que permita a los criadores, tomar decisiones de selección y hacer un uso diferencial de los mismos. La ganadería consta de distintos factores, que se deben considerar para que la misma sea exitosa y rentable, como ser la alimentación, la reproducción, la sanidad y la genética, entre otros.

La reproducción de bovinos mediante la Inseminación Artificial (IA) [2,3] es bastante sencilla y tiene muchas ventajas. Esta técnica se está aplicando desde hace bastante tiempo en el país. Hoy la Inseminación Artificial a Tiempo Fijo (IATF) es una técnica que, mediante la utilización de hormonas, permite sincronizar los celos y ovulaciones. Gracias a esto, es posible, inseminar una gran cantidad de animales en un corto período de tiempo [4].

[1] http://laslilas.com
[2] https://www.angus.org.ar/

Una de las herramientas utilizadas, por los ganaderos, para realizar las evaluaciones genéticas de los toros reproductores es la *Diferencia Esperada entre Progenie* (DEP), en base a estos valores, los productores pueden tomar decisiones de selección en base a información objetiva [5]. Los DEP anticipan cómo será el comportamiento promedio de las futuras crías de un toro en comparación con las que producirán el resto de los padres. Por el lado de las madres, a partir de esta información, la selección de los reproductores a utilizar como padres pasa a ser una de las más importantes decisiones de manejo que tiene el productor, permitiéndole seleccionar aquellos animales acordes a sus propios objetivos, su medio ambiente, su sistema de producción, e ir logrando avances genéticos que son acumulativos dentro del rodeo.

Seguidamente se detallan brevemente las descripciones de las siglas que componen los DEP's. Estas fueron extraídas del Anuario Las Lilas 2017 [6]:

- PN (Peso al nacer): Expresado en kilos, indica las diferencias genéticas para el PN de las crías de un padre determinado.
- PD (Peso al destete): Expresado en kilos. Indica el mérito genético de un reproductor en transmitir potencial de crecimiento directo a sus crías hasta el momento del destete.
- CM (Combinado materno): Esta variable combina el peso al destete y la aptitud materna en un solo valor.
- CE (Circunferencia escrotal): Expresada en centímetros y ajustada por edad de vida, es un indicador indirecto de la fertilidad de los rodeos.
- PF (Peso final): Expresado en kilos, indica la aptitud que tiene un reproductor en transmitir a su progenie capacidad de crecimiento post-destete.
- AM (Aptitud materna): Predictor de la producción lechera y aptitud materna que transmite un toro a sus hijas.
- AOB (Área del Ojo de Bife): Es la altura de la 12a costilla. Es un indicador del peso total y rendimiento de cortes despostados de la res.
- GD (Grasa dorsal): Expresada en milímetros, el espesor de grasa dorsal a la altura de la 12a costilla es un predictor genético de la precocidad y facilidad de terminación de las reses.
- MAR (Grado de Marmoreo): Es un indicador del porcentaje de grasa intramuscular del músculo dorsal largo.

Este trabajo se realiza bajo la hipótesis que si se aplica una metodología para realizar Minería de Datos (MD) a partir de los datos del material genético de los progenitores machos, más ciertos datos de las hembras, como la edad, el historial de partos, etc., se puede construir un modelo predictivo que mejor determine la etiqueta Peso al Nacer (Ver Tabla 1). Por otro lado, mediante los algoritmos No Supervisados, se intenta demostrar las relaciones existentes entre las variables, que también tengan mayor injerencia en el PN de los terneros.

**Tabla 1.** Descripción de las variables del conjunto de datos.

| Nomenclatura | Tipo de dato | Descripción |
|---|---|---|
| Peso Adulto | *Numérico* | Peso real del padre |
| PAN | *Numérico* | Peso al nacer |
| PAD | *Numérico* | Peso al destete |
| PAF | *Numérico* | Peso final |
| Circ. Escrotal | *Numérico* | Circunferencia escrotal |
| FRAME | *Numérico* | Altura del animal |
| Certificado | *Numérico* | Edad promedio de los vientres primerizos |
| PN | *Numérico* | DEP del Toro Progenitor |
| PD | *Numérico* | |
| AM | *Numérico* | |
| CM | *Numérico* | |
| PF | *Numérico* | |
| CE | *Numérico* | |
| AOB | *Numérico* | |
| GD | *Numérico* | |
| MAR | *Numérico* | |
| Peso al nacer | *Numérico* | Datos de la vaca |
| Peso al destete | *Numérico* | |
| Cantidad nacimientos | *Numérico* | |
| Cantidad abortos | *Numérico* | |
| Baja antes del destete | *Numérico* | |
| Edad en meses | *Numérico* | |
| Certificado | *Numérico* | DEP del Toro Progenitor de la vaca |
| CE del padre | *Numérico* | |
| PN | *Numérico* | |
| PD | *Numérico* | |
| AM | *Numérico* | |
| CM | *Numérico* | |
| PF | *Numérico* | |
| CE | *Numérico* | |
| AOB | *Numérico* | |
| GD | *Numérico* | |
| MAR | *Numérico* | |
| Peso_Nac | *Texto* | Atributo Clase |

En el primer año de la investigación se seleccionaron diferentes técnicas del tipo Supervisadas y No Supervisadas [7,8].

Las primeras son aquellas orientadas a predecir el valor de un atributo (etiqueta o clase) de un conjunto de datos, conocidos otros

atributos (atributos descriptivos). A partir de datos, cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Las segundas, también conocidas como técnicas de Clustering, agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud. De esta forma se agrupan las clases que sean similares entre sí y distintas con las otras clases.

Existen en la actualidad distintas metodologías para llevar a cabo un proceso de MD, en este trabajo se optó, siguiendo la literatura consultada, por Cross Industry Standard Process for Data Mining (CRISP-DM) [8,9]. Esta tecnología interrelaciona diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco.

Para la ejecución del modelo construido y las pruebas para realizar una comparación entre las diferentes técnicas, se utilizó el software WEKA[3] (acrónimo de Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato»). Para realizar las mismas, fue necesario la construcción de un modelo con los datos históricos de anteriores rodeos y sus resultados respecto al peso al nacer de sus crías. Para optimizar el modelo fue necesario normalizar las variables de entrada. Normalizar significa, en este caso, comprimir o extender los valores de la variable para que estén en un rango definido. Se empleó la fórmula de Normalización mínimo-máximo, la cual transforma linealmente los datos a un intervalo, para este caso, entre 0 y 1, donde el valor mínimo se escala a 0 y el máximo a 1 [10], que se define como:

$$X_{normalizada} = X - X_{min} / X_{max} - X_{min} \quad \textbf{(1)}$$

Para la evaluación de los clasificadores se utilizó una Matriz de Confusión (MC) y el análisis o curvas ROC (acrónimo de Receiver Operating Characteristic) [8], estos se encuentran dentro del software WEKA para cada clasificador. Con estas herramientas se evalúa entre otras cosas:

- La sensibilidad indica la capacidad de nuestro clasificador para dar como casos positivos los casos que realmente lo son; proporción de PN altos correctamente identificados.
- La especificidad indica la capacidad de nuestro estimador para dar como casos negativos los casos que realmente lo sean; por ejemplo: proporción de PN bajos correctamente identificados como tal.
- Este trabajo tiene como objetivo principal generar conocimiento especializado en el área de Minería de Datos en lo referente a un programa de mejoramiento genético. Específicamente, esta línea se centra principalmente en el estudio de dos ejes: los algoritmos Supervisados y No Supervisados.
- Además, como ya se mencionó, los datos de entrada, para realizar las pruebas contaron con los valores genéticos de sus padres/abuelos. El mayor obstáculo que se presentó es la cantidad y calidad de los mismos, si bien es posible encontrar tendencias y verificar ciertos resultados basados en el conocimiento del dominio, aún es necesario trabajar en tales datos, siendo lo deseado obtener nuevas cifras de establecimientos ganaderos similares.
- En algunos casos, incrementar el número de variables, mejora el rendimiento de los clasificadores, se pretende agregar variables tales como: el tipo de alimentación de los animales, el factor climático, etc.
- Se estudiarán nuevos algoritmos y con varias configuraciones para comprobar cual tiene la mayor capacidad de exactitud en su predicción.
- Por último, utilizar diferente herramienta computacional, por ejemplo probar con software más sofisticados como Matlab[1], SPSS[2], etc.

## 2. LÍNEAS DE INVESTIGACIÓN y DESARROLLO

Este trabajo tiene como objetivo principal generar conocimiento especializado en el área de Minería de Datos en lo referente a un

[3] www.cs.waikato.ac.nz/~ml/weka/

programa de mejoramiento genético. Específicamente, esta línea se centra principalmente en el estudio de dos ejes: los algoritmos Supervisados y No Supervisados.

Además, como ya se mencionó, los datos de entrada, para realizar las pruebas contaron con los valores genéticos de sus padres/abuelos. El mayor obstáculo que se presentó es la cantidad y calidad de los mismos, si bien es posible encontrar tendencias y verificar ciertos resultados basados en el conocimiento del dominio, aún es necesario trabajar en tales datos, siendo lo deseado obtener nuevas cifras de establecimientos ganaderos similares.

En algunos casos, incrementar el número de variables, mejora el rendimiento de los clasificadores, se pretende agregar variables tales como: el tipo de alimentación de los animales, el factor climático, etc.

Se estudiarán nuevos algoritmos y con varias configuraciones para comprobar cual tiene la mayor capacidad de exactitud en su predicción.

Por último, utilizar diferente herramienta computacional, por ejemplo probar con software más sofisticados como Matlab[1], SPSS[2], etc.

## 3. RESULTADOS OBTENIDOS Y ESPERADOS

Este grupo de investigación viene trabajando en proyectos PROINCE en años anteriores, también asociados a la misma temática:

- PROINCE C176 (2015-2016). "*Análisis Comparativo de Modelos de Clasificación de Minería de Datos (Data Mining). Su Aplicación en la Predicción de Perfiles de Alumnos en Riesgo de Deserción"*.
- PROINCE C199 (2017-2018). "*Modelos de Minería de Datos para el Diagnóstico Pre_ coz de Enfermedades Neurodegenerativas"*.
- PROINCE C205 (2017-2018). *"Uso de Minería de Datos para Acelerar la Recuperación de Documentos"*.

Los resultados de estas investigaciones dieron lugar a varias publicaciones [11-15].

Tal como quedó expuesto, se utilizaron dos tipos distintos de técnicas de MD. Se presentaron 2 trabajos en el año 2019. En uno se realizó una comparación en cuanto al desempeño de tres algoritmos del tipo Supervisado [16]:

- Árboles de Decisión (AD).
- Redes Neuronales Artificiales (RNA).
- Máquinas de Soporte Vectorial (MSV o SVM, del inglés Support Vector Machine)

En el otro trabajo presentado, se compararon cuatro algoritmos No Supervisados [17]:

- Expectation Maximization (EM).
- FarthestFirst.
- Simple K-Means.
- Mapas Auto Organizados (Redes SOM).

En el primero, se encontró que el modelo propuesto conserva una precisión (proporción de instancias clasificadas correctamente) aceptable en el caso del algoritmo AD, con un porcentaje levemente superior al 70%.

Para los algoritmos No supervisados, fue necesario un subconjunto de atributos menor del conjunto total inicial, que incluya aquellos relevantes para la tarea de agrupamiento. Se llegó a la conclusión que los datos relevantes para agrupar a los terneros según su PN involucra principalmente las características de su Madre y del Abuelo Materno.

En el segundo año se espera que, los criterios mencionados anteriormente, logrados de forma automática por los algoritmos, se puedan comparar con la estimación de los expertos en la realidad de los terneros a nacer este año. Es importante notar que la clasificación puede diferir. Sobre todo, teniendo en cuenta, la dificultad de calcular el peso del recién nacido. Por último, se están realizando con la Sociedad Rural de Chascomús y la Asociación Argentina de Angus, esto posibilitaría la obtención de datos de mejor calidad y de nuevos establecimientos ganaderos.

## 4. FORMACIÓN DE RECURSOS HUMANOS

Parte del grupo de desarrollo del proyecto trabaja desde el año 2015 en diversas áreas relacionadas con la Minería de Datos.

Actualmente forman parte del equipo, además de docentes de la UNLaM, una Ingeniera Agrónoma que trabaja en el SENASA y dos alumnos becarios de investigación.

Por otra parte, los docentes-investigadores

que integran el proyecto dictan clases en la cátedra de Inteligencia de Negocio de la carrera Licenciatura en Gestión de Tecnología y en la cátedra Base de Datos y Data Mining y Data Warehouse de la carrera de Ingeniería Informática. Se prevé, además, la capacitación y formación de recursos humanos, a través de cursos de actualización y posgrado en el área de estudio; la transferencia de conocimiento y resultados; y la posibilidad de brindar charlas informativas del desarrollo e implementación del proyecto a distintas instituciones del sectores ganadero, como la Sociedad Rural de Chascomús y la Asociación Argentina de Angus.

## 5. BIBLIOGRAFÍA


1. Firpo Brenta, L. y otros. (2012). Selección genética y mejoramiento animal. Disponible en: http://www.produccion-animal.com.ar /genética_seleccion_cruzamientos/bovinos_en_general/24-Seleccion_genetica.pdf. Último acceso: 06/09/2019.
2. Agrocor. (2011). Inseminación artificial en bovinos Curso Teórico Práctico de Inseminación Artificial en Bovinos. Disponible en: https://www.engormix.com /ganaderia-carne/articulos/inseminacion-artificial-en-bovinos-t26957.htm. Último acceso: 06/09/2019.
3. Díaz, P. Fonseca, V. Martínez P. y Rey A. (2003). Inseminación Artificial en bovinos. Biblioteca Digita, U. de Chile. Disponible en: www.biblioteca.org.ar/libros/8913.pdf. Último acceso: 06/02/2020.
4. Sommantico, S. (2018). Inseminación Arti_ ficial a Tiempo Fijo: la tecnología de la que se habla mucho y se usa poco. Disponible en: https://www.infocampo.com.ar/inseminacion-artificial-a-tiempo-fijo-la-tecnologia-de-la-que-se-habla-mucho-y-se-usa-poco/ Último acceso: 26/02/2020.
5. Guitou H. y Monti A. (1998). Interpretación y uso correcto de los DEPs como herramienta de selección. INTA Castelar. Disponible en: https://es.scribd.com/document/337947981/20-Interpretacion-Deps. Último acceso: 26/02/2020.
6. Cómo interpretar la evaluación genética. Anuario Las Lilas 2018-2019. Cabaña Las Lilas. Centro de Genética. Pág. 107. Disponible en: http://laslilas.com/pdf/Anuario-Genetica-2017.pdf. Último acceso: 26/02/2020.
7. Perez Lopez, C. y otros. (2007). Minería de datos. Técnicas y herramientas ISBN: 9788497324922. Ed. Paraninfo Cengage L. Madrid. España.
8. Hernández Orallo, J. y otros. (2004). Introdu_ cción a la minería de datos". Pearson. Edición: I.
9. Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. (2005). U. de Alcalá, Madrid Disponible en: http://www.sc.ehu.es /jiwdocoj/remis/docs/GarreAdis05.pdf. Ultimo acceso: 26/02/2020.
10. Han Jiawei. Data Mining: Concepts and Techniques. 3ra. Edición. (2011). ISBN 978-0-12-381479-1.
11. Predicción del riesgo de abandono universitario utilizando métodos supervisados. (2016) IPECyT 2016. F. R. B. Blanca.
12. "Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores". (2016). el CONAIISI 2016. Salta. Argentina.
13. "Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz". Presentado CONAIISI 2017. Santa Fe. Argentina.
14. "Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R". Presentado en CACIC 2018. U. N. del Centro, Tandil.
15. "Selection of voice parameters for Parkinson´s disease prediction from collected mobile data". (2019) XXII Symposium on Image, Signal Processing and Art. Vision. Bucaramanga, Colombia.
16 *"Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados"*. Trabajo presentado y no publicado en JAIIO 2019. UNSa.
17 *"Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados"*. Trabajo presentado en CONAIISI 2019. UNLaM. Buenos Aires. Argentina.